# Pre-Analysis, Data Cleaning, Data Transformation & EDA Notes

## Pre-Analysis

The original dataset was extracted from the US FDA webpage on the 13th February 2025. It consists on 3 files in txt format:

* exclusivity\_raw.txt
* patents\_raw.txt
* products\_raw.txt

Being my first health data project, it seems an interesting dataset to start with as it feels manageable and sufficiently challenging. I hope this analysis brings impactful insights on patents and exclusivities associated with FDA-approved drugs.

In order to come up with insightful questions to answer when analysing the dataset is crucial to understand all data fields. For that reason, we have revised the metadata provided by the FDA (see Metadata\_FDA\_Orange\_Book.docx) where we have made some notes and pointed out interesting questions to answer:

Products table:

* *Mode/median/mean number of API in the composition of FDA-approved drugs. Is it different between innovators and generics? Is it different between Rx and OTC?*
* *What percentage represent combination drugs (multiple APIs in their composition) with respect to the total FDA-Approved product? Is it different between innovators and generics? Is it different between Rx and OTC?*
* *Which are the most popular APIs in FDA-approved products’ composition? Is it different between innovators and generics? Is it different between Rx and OTC?*
* *Which are the most common dosage forms and administration routes in FDA-approved drugs? Is it different between innovators and generics?* *Is it different between Rx and OTC?*
* *Which are the firms owning the highest quantity of FDA-approved products (innovators/generics/Rx/OTC/)? And by Dosage Form? And by Route of administration?*
* *What percentage represent innovators and generics with respect to the total number of FDA-approved products? And within prescription drugs? Or within OTC? Which percentage represents Rx and OTC within innovators? And within generics?*
* *How many different new drug applications (NDA or ANDAs) are registered at the FDA orange book at this moment*?
* *How many different products (product numbers) are normally included in a new drug application that has been approved by the FDA (mode, mean, median)? Is it different for innovators and generics?* *Is it different between Rx and OTC?*
* *What percentage of FDA-approved generics are considered therapeutically equivalent to their reference listed drug (RLD)? Which percentage is not?*
* *Which has been the FDA approval tendency over the past years?*
* *What percentage of RLDs are considered RS by the FDA?*
* *Which firms have the highest numbers of FDA discontinued drugs?*

Patents table:

* *How many FDA-approved products (drugs) are registered at the FDA Orange book at this moment and have an associated patent? How many different products (product numbers) are normally included in a new drug application that has been approved by the FDA and have an associated patent (mode, mean, median)? Is it different for innovators and generics? Is it different between Rx and OTC?*
* *Which is the total number of patents from FDA-approved drugs?*
* *Which has been the patent expiration tendency over the years?*
* *In what percentage of patents does the applicant claim the drug substance?*
* *In what percentage of patents does the applicant claim the drug product?*
* *Mean/median/mode number of different uses/indications that are commonly claimed in a patent (use patent use codes).*
* *What percentage of patents have been requested to be to delisted by the applicant? Which applicants have requested to delist patents the most (take into account the total number of patents per applicant)?*
* *Mean/median/mode patent duration. Which has been the patent submission tendency throughout the years?*

Exclusivities table:

* *How many applications are associated with innovators / generics products?*
* *How many different applications are registered at the FDA orange book?*
* *How many different products (product numbers) are normally included in NDA/ANDA (mode/median/mean)? It is different for innovators or generics?*
* *Which are the most popular post-approval exclusivities granted by the FDA?*
* *Do certain exclusivities tend to be granted together? (e.g., do some applications have multiple exclusivity types?)*
* *Which exclusivity codes are found only in innovator applications?*
* *Which exclusivity codes are found only in generic applications?*
* *How many exclusivity periods expire each year in the dataset?*
* *What percentage of applications receive exclusivity? (for this I need those applications not appearing at the exclusivities table)*

## Data Cleaning and Transformation

The next step is to dive into the dataset. To explore the dataset we have opted to use pandas, a python library very useful for data manipulation, cleaning and transformation. See:

* pre\_Analysis\_cleaning\_transformation\_ex.ipynb
* pre\_Analysis\_cleaning\_transformation\_pat.ipynb
* pre\_Analysis\_cleaning\_transformation\_prod.ipynb

## EDA (Exploratory Data Analysis)

Once we have cleaned and transformed our dataset, we will perform univariate and multivariate analysis on our columns. For this aim, we will choose the most popular data visualization libraries in python: matplotlib and seaborn.